Biometrika (1989), **76**, 1, *pp.* 117–23
Printed in Great Britain

Alternative variance and efficiency calculations for the case-cohort design

BY SHOLOM WACHOLDER AND MITCHELL H. GAIL
Biostatistics Branch, National Cancer Institute, Rockville, Maryland 20892, U.S.A.

DAVID PEE
Information Management Services, Rockville, Maryland 20852, U.S.A.

AND RON BROOKMEYER
Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

SUMMARY

We define a bootstrap sampling plan for the case-cohort design and present bootstrap variance estimates and confidence intervals for the log relative hazard. The coverages of these confidence intervals are found to be near nominal levels in simulations. A simple null variance calculation, based on a superpopulation model, is shown to provide good estimates of the power and efficiency of the case-cohort design.

Some key words: Bootstrap; Cohort study; Epidemiologic study; Proportional hazards model.

1. INTRODUCTION

The case-cohort design (Prentice, 1986) is economical for studying large cohorts because covariates need not be measured on all members of the cohort. All members of the cohort are followed until 'death' or censoring, but covariate information is only obtained for those who die and for all members of a subcohort, $\tilde{C}$, that was randomly selected from the entire cohort at the outset; see Table 1. However, estimation of the variance of $\tilde{\beta}$, the estimate of the regression coefficient β in a model with hazards proportional to $\exp(\beta Z)$, is complex because contributions to the score statistic are correlated (Prentice, 1986). For ease of exposition, we take Z to be scalar.

The objective of this paper is to study the operating characteristics of two alternative variance estimators for $\tilde{\beta}$. The first is based on a superpopulation model and is valid under the null hypothesis $\beta = 0$, provided censoring is independent of Z, as we assume. This superpopulation variance estimator, $\tilde{V}_{SP}$, is attractive because it yields simple

Table 1. *Numbers of subjects observed in a case-cohort experiment*

	Number of deaths	Number censored alive	Total
Subcohort	Q	$M-Q$	$M=\rho N$
Others	$D-Q=\Sigma\Delta_j$	—	$(1-\rho)N$
Total cohort	D	$N-D$	N

Covariate information not available for $N-D-(M-Q)$ subjects who neither die nor are members of subcohort. Quantity $\Sigma\Delta_j$ is number of deaths among 'others'.

estimates of local power and relative efficiency and provides insight into the covariance structure among contributions to the score. The second approach is based on a modified bootstrap.

We assume that all study subjects are at risk from the beginning of the study, as in the clinical trial described by Prentice (1986), so that successive subcohort risk sets are subsets of previous subcohort risk sets.

2. VARIANCE ESTIMATES

2·1. *Variance estimates based on superpopulation model when* $\beta = 0$

The kth ordered death time in the full cohort is called t_k, and we let n_k be the number of members of the subcohort still at risk at t_k. We call this risk set within the subcohort $\tilde{C}_k$. The indicator variable Δ_k is set to 1 if the kth death is outside the subcohort and 0 otherwise. The covariate value corresponding to the kth death is Z_k, and we let Z_{k+} denote the sum of the covariates over all members in $\tilde{C}_k$.

To compute $\tilde{V}_{SP}$, we assume that $\tilde{C}$ is a random sample from an infinite population with $E(Z) = \mu$ and var $(Z) = \sigma^2$. Assuming the null hypothesis $\beta = 0$, and that censoring is independent of Z, we can treat $\tilde{C}_k$ as a random sample drawn without replacement from $\tilde{C}$, and hence from the original superpopulation. We estimate σ^2 as the sample variance of the Z values in the subcohort. All the covariances below are conditional on $\{n_k\}$ and can be used as estimates of unconditional covariances since the related scores have conditional expectation zero. The variance of the score

$$U_k = Z_k - (Z_{k+} + \Delta_k Z_k)(n_k + \Delta_k)^{-1} = (n_k Z_k - Z_{k+})(n_k + \Delta_k)^{-1}$$

is

$$\text{var}\,(U_k) = \sigma^2(n_k - 1 + \Delta_k)(n_k + \Delta_k)^{-1}. \tag{1}$$

To compute cov (U_k, U_j) with $t_j > t_k$, note that $\tilde{C}_j$ is drawn without replacement from $\tilde{C}_k$. It follows that cov $(Z_k, Z_j) =$ cov $(Z_k, Z_{j+}) \equiv 0$, but cov $(Z_j, Z_{k+}) = \sigma^2(1 - \Delta_j)$ because the subject who died at t_j was in $\tilde{C}_k$ only if he was in the subcohort. Likewise cov $(Z_{k+}, Z_{j+}) = \sigma^2 n_j$ since all n_j members of $\tilde{C}_j$ were in $\tilde{C}_k$. From these relations we obtain

$$\text{cov}\,(U_k, U_j) = \sigma^2 n_j \Delta_j \{(n_k + \Delta_k)(n_j + \Delta_j)\}^{-1}. \tag{2}$$

Thus a positive covariance in (2) arises primarily from overlapping summands in Z_{k+} and Z_{j+}, but the covariance (2) vanishes when $\Delta_j = 0$, because then

$$E(U_k U_j) = E(U_k) E(U_j \mid Z_k, Z_{k+}) = E(U_k) E\{n_k^{-1}(Z_{k+} - Z_{k+})\} = 0.$$

Note the conditioning differs from that in a similar argument of Prentice (1986). The estimate of var $(\tilde{\beta})$, namely $\tilde{V}_{SP}$, is obtained from (2) and (3) and is valid only when $\beta = 0$. We shall exploit it to study efficiency and to compute sample size for β near 0, but do not advocate its use for analysis.

2·2. *Variance estimates based on the bootstrap*

Standard bootstrap methods (Efron, 1982; Efron & Tibshirani, 1986) cannot be used because covariates are not available on all subjects in the cohort. However, the following bootstrap sampling scheme yields good results. A more complex bootstrap procedure that allowed the number of events in the subcohort to vary yielded similar results.

For $b = 1, \ldots, B$ proceed as follows.

(i) Obtain a bootstrap sample of D cases by sampling from the original D cases with replacement. Assign the first Q sampled cases to the subcohort; the remaining $D - Q$ fall outside the subcohort.

(ii) Obtain a bootstrap sample of $M - Q$ noncases by sampling with replacement from the original $M - Q$ noncases in the subcohort. The $M - Q$ subjects will constitute the noncases in the bootstrap subcohort.

(iii) The resulting bootstrap sample has the same numbers in each cell of Table 1 as the original case-cohort sample. Calculate the pseudo-maximum likelihood estimate $\tilde{\beta}_b$ just as for the original sample estimate $\tilde{\beta}$ from the pseudo-score equation (6) of Prentice (1986). With extreme designs and discrete Z, all cohort members and cases may be concordant for exposure, in which case there is no statistical information about β. In such cases, we discard the current bootstrap sample and replace it with another, just as we would ignore such results if they had occurred in the original case-cohort experiment. For dichotomous exposures, the quantity $\tilde{\beta}^*$ that maximizes the pseudolikelihood of Prentice (1986) can be infinite, as when all D persons who died were exposed. We therefore define our estimate $\tilde{\beta}$ as $\{\text{sgn}\,(\tilde{\beta}^*)\}\{\min\,(|\tilde{\beta}^*|, \beta^*)\}$, where the truncation point

$$\beta^* = \log\,[(D+\tfrac{1}{2})\{(1-\hat{p}_e)(M-Q)+\tfrac{1}{2}\}(\tfrac{1}{2})^{-1}\{\hat{p}_e(M-Q)+\tfrac{1}{2}\}^{-1}],$$

and where $\hat{p}_e$ is the proportion of subcohort members who are exposed. The truncation point β^* is the logarithm of the odds ratio in a classification on survival and exposure in which $\frac{1}{2}$ has been added to each cell and in which all persons who died were exposed. These truncations rarely come into play in bootstrapping, and almost never are required in the original partial likelihood or case-cohort estimates of β. For consistency the same truncation is used in all cases.

Obtain the bootstrap variance estimator (Efron, 1982) from $\tilde{V}_B = (B-1)^{-1}\Sigma(\tilde{\beta}_b - \bar{\beta})^2$, where $\bar{\beta}$ is the average of the B bootstrap estimates. From the preceding comment, $\tilde{V}_B$ may be regarded as the variance estimate for the truncated case-cohort estimator.

3. Simple sample size and efficiency calculations for small β based on the superpopulation model

Self & Prentice (1988) computed the asymptotic relative efficiency of the case-cohort design, relative to the full cohort design, for a fixed binary covariate and arbitrary β. However, they restricted calculations to the case of simultaneous entry into the cohort, no losses to follow-up, and final analysis at a fixed time t. Efficiency calculations based on $\tilde{V}_{SP}$ are valid only for β near 0 but do allow for competing risks and other forms of loss to follow-up as well as for a period of accrual.

Our approach is to use (1) and (2) by replacing n_k and n_j by their expected values at the times t_k and t_j, based on knowledge of the accrual plan, competing risks, and failure rates (Rubinstein, Gail & Santner, 1981). The procedure is as follows. First, solve equation (A2) of Rubinstein et al. (1981) to obtain T_j, the expected time required until death j ($j = 1, \ldots, D$). Determine the expected number at risk at T_j, namely, $N_j = NS(T_j)C(T_j)$, where S and C are the assumed survivor functions for death and censoring, respectively, and N is the original cohort size. Calculate $\Sigma\,\text{var}\,(U_j)$ and $\Sigma\,\text{cov}\,(U_j, U_k)$ from (1) and (2) by replacing n_k and n_j by ρN_k and ρN_j, respectively, and Δ_j by its expectation, $1-\rho$.

Then obtain the estimate (Prentice, 1986)

$$\tilde{V}_{SP} = \{\Sigma \tilde{v}(U_j) + 2\sum_{k<j} \tilde{c}(U_k, U_j)\}\{\Sigma \tilde{v}(U_j)\}^{-2}, \quad (3)$$

where $\tilde{v}$, $\tilde{c}$ are estimated variances and covariances. For the full cohort, the Cox (1972) estimate $\hat{\beta}$ has variance estimator $\hat{V}_c$ obtained by setting $\rho = 1$ in this procedure. The asymptotic relative efficiency of the case-cohort to the full cohort design is estimated as $\hat{V}_C / \tilde{V}_{SP}$. The variance estimator for $\tilde{\beta}$ given by Prentice (1986) is denoted by $\tilde{V}_P$.

4. SIMULATION STUDIES

4·1. *Simulation methods*

Several designs were simulated to study $\tilde{V}_P$, $\tilde{V}_{SP}$, $\tilde{V}_B$ and the coverage of their associated confidence intervals. For each design, 1000 independent cohorts, each of size $N = 1000$ were simulated, except for two cases in which $N = 5000$ was used. Exactly Np_e subjects with $Z = 1$, exposed, and $N(1-p_e)$ subjects with $Z = 0$, unexposed, were in each cohort. Independent exponential failure times with hazards 1 and e^{β} were generated for the exposed and unexposed groups, respectively. Independent, exponentially distributed failure times for competing risks were also generated with hazards ϕ_0 and ϕ_1 respectively for those with $Z = 0$ and $Z = 1$. Independent cohort accrual times were uniform on $[0, \tau)$, and observation continued beyond τ to time T, when the experiment ended. Thus, the maximum duration of follow-up was uniformly distributed on $[T-\tau, T]$. We specified the required total number of expected events and the ratio of accrual time to total experimental time, τ/T. Then T was determined following Rubinstein et al. (1981). To simulate the case-cohort experiment, we randomly selected a subcohort of size $M = \rho N$ without replacement.

For each simulated cohort, we obtained $\hat{\beta}$, $\tilde{\beta}$ and four estimates of var $(\tilde{\beta})$, namely $\tilde{V}_P$, $\tilde{V}_{SP}$ and $\tilde{V}_B$ for $B = 50$ and $B = 200$. Ninety-five percent confidence intervals from the Cox estimator were computed as $\hat{\beta} \pm 1{\cdot}96\hat{V}_c^{\frac{1}{2}}$. Similar confidence intervals were computed from $\tilde{\beta}$. We also obtained a confidence interval from the 2·5% and 97·5% percentiles of the $B = 200$ bootstrap replications.

Random numbers were obtained from the International Mathematical and Statistical Libraries (1984) subroutines GGUBS and GGEXN. A typical run with 1000 simulations took about 1000 seconds of central processor time on a Cray X-MP 24 supercomputer.

4·2. *Simulation results*

The coverages are near the nominal 95% level for all procedures shown in Table 2, and $\tilde{V}_{50}$ performs as well as $\tilde{V}_{200}$. It is noteworthy that the percentile method works well with only $B = 200$ replications (Efron, 1987). The variance is not greatly affected by cohort size, the accrual pattern, data not shown, or the strength of the competing risk, but, as theory predicts, the variance does tend to decrease with increasing numbers of expected deaths, and with subcohort size. The variance is lower for p_e values of 0·1 and 0·9 than for $p_e = 0{\cdot}5$. The ratios of the empirical variance of $\tilde{\beta}$ to the average values of $\tilde{V}_P$, $\tilde{V}_{SP}$ and $\tilde{V}_B$ are within 10% of 1·0 except for $\tilde{V}_B$ with $p_e = 0{\cdot}1$ and 0·9. For these designs, only 10 exposed and unexposed cases, respectively, are expected in the entire cohort, and thus only 2 in the subcohort. The bias, not shown, in estimating β was always close to zero except when $p_e \neq 0{\cdot}5$. Even with $p_e = 0{\cdot}1$ or 0·9, however, the coverage of

Table 2. *Simulations of coverage of confidence intervals*

D	Special condition†	Full cohort $\hat{V}_C$	Case-cohort $\tilde{V}_P$	$\tilde{V}_{SP}$	$\tilde{V}_{50}$	$\tilde{V}_{200}$	%
			(a) *Null case*				
30	$M=2D$	946	951	944	952	952	945
		0·143	0·211	0·209	0·227	0·226	
		1·05	0·99	1·00	0·92	0·93	
50	$M=2D$	948	949	947	957	957	944
		0·083	0·124	0·124	0·130	0·130	
		1·07	1·00	1·00	0·95	0·95	
50	$\phi_0=5, \phi_1=10$	961	956	954	955	960	963
		0·084	0·127	0·126	0·134	0·134	
		0·99	0·99	1·00	0·94	0·94	
100	$p_e=0·9$	962	951	942	962	965‡	948
		0·123	0·185	0·180	0·217	0·220	
		1·04	1·02	1·04	0·87	0·85	
100	$p_e=0·5$	959	957	958	952	957	954
		0·041	0·062	0·062	0·063	0·063	
		0·98	0·92	0·92	0·90	0·90	
100	$p_e=0·1$	955	952	939	964	973‡	943
		0·123	0·184	0·178	0·242	0·239	
		1·06	1·06	1·10	0·81	0·81	
			(b) *Alternative case*				
30	$\beta=\log 3$	962	947		960	958	934‡
		0·207	0·271		0·310	0·312	
		1·13	1·10		0·96	0·96	
50	$M=2D$	964	955		952	958	949
		0·094	0·133		0·139	0·141	
		0·96	0·98		0·94	0·93	
50	$\phi_0=5, \phi_1=10$	950	950		951	956	951
		0·098	0·139		0·147	0·148	
		1·10	1·05		0·99	0·98	
100	$p_e=0·9$	965	962		949	950	932‡
		0·245	0·300		0·353	0·356	
		1·07	1·07		0·90	0·90	
100	$p_e=0·5$	957	960		958	961	957
		0·045	0·065		0·066	0·067	
		0·99	1·01		0·99	9·98	
100	$p_e=0·1$	947	958		966‡	969‡	955
		0·073	0·134		0·148	0·148	
		1·01	1·01		0·91	0·91	

Top number, coverage of nominal 95% confidence interval; middle number, average value of corresponding variance estimate for $\tilde{\beta}$; bottom number, ratio of variance estimate $s^2=\Sigma(\tilde{\beta}-\beta)^2/999$ to average estimated variance. V_C, $\tilde{V}_P$, $\tilde{V}_{SP}$, $\tilde{V}_{50}$ and $\tilde{V}_{200}$, procedures based on variance estimates; %, bootstrap percentile confidence interval.

† Unless otherwise noted, all simulations have cohort size $N=1000$, $\beta=0$ for null cases, $\beta=\log 2$ for alternatives, subcohort size $M=2D$, accrual time $\tau=T/2$, proportion exposed $p_e=0·5$, and hazards of competing risk $\phi_0=\phi_1=5$, where ϕ_0 corresponds to $X=0$ and ϕ_1 to $X=1$.

‡ Estimated coverage is outside interval (936, 964), which should contain 95% of cases provided coverage equals nominal 95%.

these confidence intervals was near nominal levels. Coverages were less than nominal for confidence intervals based on $\tilde{V}_{SP}$ with $\beta \neq 0$, data not shown. Failure to account for covariances in (3) leads to underestimates of variance of 30% in typical cases.

The quantity $\tilde{V}_{SP}$ provides a good guide to observed power for $p_e = 0{\cdot}5$, even for the nonlocal alternative $\beta = \log 3$; see Table 3. Power increases with increasing numbers of deaths and with subcohort size M. With $M = 8D = 400$, the subcohort design has nearly the same power as the full cohort design. For fixed $D = 50$, $\beta = \log 2$, $p_e = 0{\cdot}5$ and $M = 2$, other aspects of the design characterized by τ, ϕ_0, ϕ_1 and N have very little impact on power as predicted by the superpopulation model and as found empirically. The use of $\tilde{V}_{SP}$ leads to overestimates of power for $\beta = \log 2$ when $p_e = 0{\cdot}1$ and to underestimates when $p_e = 0{\cdot}9$, both for the full cohort and for the case-cohort analyses. This is probably because, with $p_e = 0{\cdot}1$, some risk sets become exhausted, or nearly exhausted, of exposed individuals as the trial proceeds and provide little information on β, whereas, for $p_e = 0{\cdot}9$, loss of exposed individuals produces more nearly balanced numbers of exposed and unexposed members in later risk sets, and these balanced risk sets provide more information about β.

Calculations of $\tilde{V}_{SP}$ also provide a good guide to the ratio of the empirical variances of $\hat{\beta}$ and $\tilde{\beta}$; see Table 3. These ratios are based on independent sets of 1000 simulations with $\beta = 0$ which are distinct from the other simulations used to study power in Table 3. Note that, with $M = 8D = 400$, the efficiency of the case-cohort design is calculated as 0·92, in good agreement with the observed variance ratio 0·94, and slightly higher than would be predicted from the formula $8/(8+1) = 0{\cdot}89$, which is the ratio of the variance for a case-control study with 8 controls per case compared to the variance with infinite controls per case (Ury, 1975).

Table 3. *Validity of superpopulation variance as guide to power and efficiency*

		Full cohort power		Case-cohort power				Variance ratios‡	
D	Special conditions†	ψ	$\hat{\psi}_C$	ψ	$\hat{\psi}_P$	$\hat{\psi}_{200}$	$\hat{\psi}_{\%}$	ρ	$\hat{\rho}$
30	$\beta = \log 3$	0·905	0·886	0·779	0·744	0·702	0·763	0·66	0·72
50	$M = D$	0·782	0·770	0·515	0·493	0·479	0·502	0·48	0·52
50	$M = 2D$	0·782	0·755	0·628	0·585	0·575	0·581	0·66	0·72
50	$M = 4D$	0·782	0·769	0·706	0·677	0·670	0·678	0·81	0·80
50	$M = 8D$	0·782	0·776	0·752	0·761	0·744	0·761	0·92	0·94
50	$\tau = 0$	0·789	0·715	0·637	0·577	0·562	0·590	0·66	0·69
50	$\tau = T$	0·782	0·773	0·594	0·610	0·582	0·604	0·60	0·64
50	$\phi_0 = \phi_1 = 0$	0·782	0·762	0·640	0·605	0·591	0·625	0·68	0·70
50	$\phi_0 = \phi_1 = 15$	0·782	0·770	0·591	0·598	0·570	0·610	0·60	0·66
50	$N = 5000$	0·782	0·776	0·630	0·629	0·611	0·628	0·66	0·65
100	$p_e = 0{\cdot}9$	0·664	0·471	0·510	0·393	0·274	0·471	0·65	0·68
100	$p_e = 0{\cdot}5$	0·964	0·968	0·873	0·884	0·882	0·885	0·65	0·70
100	$p_e = 0{\cdot}1$	0·664	0·803	0·510	0·613	0·582	0·585	0·65	0·67
100	$\beta = \log(1{\cdot}5)$	0·644	0·638	0·494	0·494	0·493	0·494	0·65	0·63

Quantity ψ, predicted power based on superpopulation model; $\hat{\psi}_C$, $\hat{\psi}_P$, $\hat{\psi}_{200}$ and $\hat{\psi}_{\%}$, fractions of rejections observed in 1000 simulations. One-sided 0·05 level rejection regions like $\tilde{\beta} > 1{\cdot}645\{\tilde{V}_p\}$ used, except for percentile method for which rejection occurred when 0 fell below the 5th bootstrap percentile.

† Unless otherwise noted, simulations use $\beta = \log 2$, cohort size $N = 1000$, subcohort size $M = 2D$, accrual time $\tau = T/2$, competing risk hazards $\phi_0 = 5$ and $\phi_1 = 5$, and proportion exposed $p_e = 0{\cdot}5$.

‡ Ratio ρ of variance of $\hat{\beta}$ to variance of $\tilde{\beta}$ computed theoretically from superpopulation model. Ratio $\hat{\rho}$ of corresponding empirical variances computed from independent sets of 1000 simulations with $\beta = 0$.

When we applied the methods in § 3 to a trial with simultaneous entry and no loss to follow-up, calculated efficiencies were in close agreement with those in Table 1 of Self & Prentice (1988). That table shows that the asymptotic relative efficiency increases only very slightly as β changes from 0 to $\log 2$, in accord with our empirical findings that calculations based on $\tilde{V}_{SP}$ are sufficiently accurate for power estimates.

5. Discussion

Our simulations confirm that a simple superpopulation model may be used to predict the power of the case-cohort design, and its efficiency, compared to the full cohort, in the presence of administrative censoring and competing risks, provided these act equally on the two exposure groups.

Simulations confirm that the bootstrap procedure has near nominal operating characteristics. With subcohort size 200 and 100 events, calculations of $\tilde{V}_{200}$ took about 4 times as long as $\tilde{V}_P$. The bootstrap calculations may be conveniently added to standard programs and may be useful in more complex case-cohort designs.

References

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B **34**, 187–220.

Efron, B. (1982). *The Jackknife, the Bootstrap, and other Resampling Plans*, Soc. Ind. Appl. Math. CBMS-Natl. Sci. Found., Monogr. 38.

Efron, B. (1987). Better bootstrap confidence intervals. *J. Am. Statist. Assoc.* **82**, 171–85.

Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* **1**, 54–75.

International Mathematical and Statistical Libraries (1984). *User's Manual.* Houston, Texas.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Rubinstein, L. V., Gail, M. H. & Santner, T. H. (1981). Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J. Chronic Dis.* **34**, 469–79.

Self, S. G. & Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16**, 64–81.

Ury, H. K. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics* **31**, 643–9.

[*Received January* 1988. *Revised June* 1988]